Chromosome 2 Sequence of the Human Malaria Parasite *Plasmodium falciparum*

Malcolm J. Gardner, Hervé Tettelin, Daniel J. Carucci, Leda M. Cummings, L. Aravind, Eugene V. Koonin, Shamira Shallom, Tanya Mason, Kelly Yu, Claire Fujii, James Pederson, Kun Shen, Junping Jing, Christopher Aston, Zhongwu Lai, David C. Schwartz, Mihaela Pertea, Steven Salzberg, Lixin Zhou,* Granger G. Sutton,† Rebecca Clayton, Owen White, Hamilton O. Smith,† Claire M. Fraser, Mark D. Adams,† J. Craig Venter,† Stephen L. Hoffman‡

Chromosome 2 of *Plasmodium falciparum* was sequenced; this sequence contains 947,103 base pairs and encodes 210 predicted genes. In comparison with the *Saccharomyces cerevisiae* genome, chromosome 2 has a lower gene density, introns are more frequent, and proteins are markedly enriched in nonglobular domains. A family of surface proteins, rifins, that may play a role in antigenic variation was identified. The complete sequencing of chromosome 2 has shown that sequencing of the A+T-rich *P. falciparum* genome is technically feasible.

Malaria, a disease caused by protozoan parasites of the genus *Plasmodium*, is one of the most dangerous infectious diseases affecting human populations. Approximately 300 million to 500 million people are infected annually, and 1.5 million to 2.7 million lives are lost to malaria each year, with most deaths occurring among children in sub-Saharan Africa (*1*). Of the four species that cause malaria in humans, *P. falciparum* is the greatest cause of morbidity and mortality. The resistance of the malaria parasite to drugs and the resistance of mosquitoes to insecticides have resulted in a resurgence of malaria in many parts of the world and a pressing need for vaccines and new drugs. The identification of new targets for vaccine and drug development is dependent on the expansion of our understanding of parasite biology; this understanding is hampered by the complexity of the parasite life cycle. The sequencing of the *Plasmodium* genome may circumvent many of these difficulties and rapidly increase our knowledge about these parasites.

The *P. falciparum* genome is ~30 Mb in size; has a base composition of 82% A+T; and contains 14 chromosomes, which range from 0.65 to 3.4 Mb. Chromosomes from different wild isolates exhibit extensive size polymorphism. Mapping studies have indicated that the chromosomes contain central domains that are conserved between isolates and polymorphic subtelomeric domains that contain repeated sequences. *P. falciparum* also contains two organellar genomes. The mitochondrial genome is a 5.9-kb, tandemly repeated DNA molecule; a 35-kb circular DNA molecule, which encodes genes that are usually associated with plastid genomes, is located within the apicoplast [an organelle of uncertain function in *Plasmodium* and the related parasite *Toxoplasma* (*2*)].

Chromosome 2 (GenBank accession number AE001362) was sequenced with the shotgun sequencing approach, which was previously used to sequence several microbial genomes (*3*, *4*), with modifications to compensate for the A+T richness of *P. falciparum* DNA (*5*). These modifications included the following: the extraction of DNA from agarose under high-salt conditions to prevent the DNA from melting at a high temperature, the avoidance of ultraviolet (UV) light, the use of the "vector plus insert" protocol for library construction, sequencing with dye-terminator chemistry, the use of a reduced extension temperature in polymerase chain reactions (PCRs), and the use of a transposon-insertion method for the closure of gaps that are very rich in AT. The assembly software was also modified to minimize the misassembly of A+T-rich sequences. The complete sequence included portions of both telomeres and had an average redundancy of 11-fold; colinearity of the final sequence and genomic DNA was proven with optical restriction and yeast artificial chromosome (YAC) maps.

Chromosome 2 of *P. falciparum* (clone 3D7) is 947 kb in length and has an overall base composition of 80.2% A+T. The chromosome contains a large central region that encodes single-copy genes and several duplicated genes, subtelomeric regions that contain variant antigen genes (*var*) (*6*–*8*), repetitive interspersed family (RIF)–1 elements (*9*) and other repeats, and typical eukaryotic telomeres (Fig. 1). The terminal 23-kb portions of the chromosome are noncoding and exhibit 77% identity in opposite orientations. The left and right telomeres consist of tandem repeats of the sequence TT(TC)AGGG (*10*) and total 1141 and 551 nucleotides (nt), respectively. The subtelomeric regions do not exhibit repeat oligomers until ~12 to 20 kb into the chromosome, where rep20 (*11*) (a 21-bp tandem direct repeat found exclusively in these regions) occurs 134 and 96 times in the left and right ends of the chromosome, respectively. The sequence similarity that was observed between the subtelomeric regions supports previous suggestions that recombination between chromosome ends may be one mechanism by which genetic diversity is generated. A region with centromere functions could not be identified on the basis of sequence similarity to *S. cerevisiae* or other eukaryotic centromeres (*12*). However, several regions of up to 12 kb are devoid of large open reading frames (ORFs) and might contain the centromere. Alternatively, centromeric functions may be defined by higher order DNA structures and chromatin-associated protein complexes (*13*).

Two hundred and nine protein-encoding genes and a gene for tRNA$^{\mathrm{Glu}}$ (Fig. 1 and Table 1) were predicted (*14*) on chromosome 2, giving a gene density of one gene per 4.5 kb, which is a value between that observed in yeast (one gene per 2 kb) and in *Caenorhabditis elegans* (one gene per 7 kb). Of the 209 protein-encoding genes, 43% contain at least one intron. This percentage is an estimate

M. J. Gardner, H. Tettelin, L. M. Cummings, S. Shallom, T. Mason, K. Yu, C. Fujii, J. Pederson, K. Shen, L. Zhou, G. G. Sutton, R. Clayton, O. White, H. O. Smith, C. M. Fraser, M. D. Adams, J. C. Venter, The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. D. J. Carucci and S. L. Hoffman, Malaria Program, Naval Medical Research Institute, 12300 Washington Avenue, Rockville, MD 20852, USA. L. Aravind, Department of Biology, Texas A & M University, College Station, TX 70843, USA, and National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. E. V. Koonin, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. J. Jing, C. Aston, Z. Lai, D. C. Schwartz, W. M. Keck Laboratory for Biomolecular Imaging, Department of Chemistry, New York University, New York, NY 10003, USA. M. Pertea, Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA. S. Salzberg, The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA, and Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA.

*Present address: ARIAD Pharmaceuticals, 26 Landsdowne Street, Cambridge, MA 02139, USA.
†Present address: Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA.
‡To whom correspondence should be addressed. E-mail: hoffmans@nmripo.nmri.nnmc.navy.mil

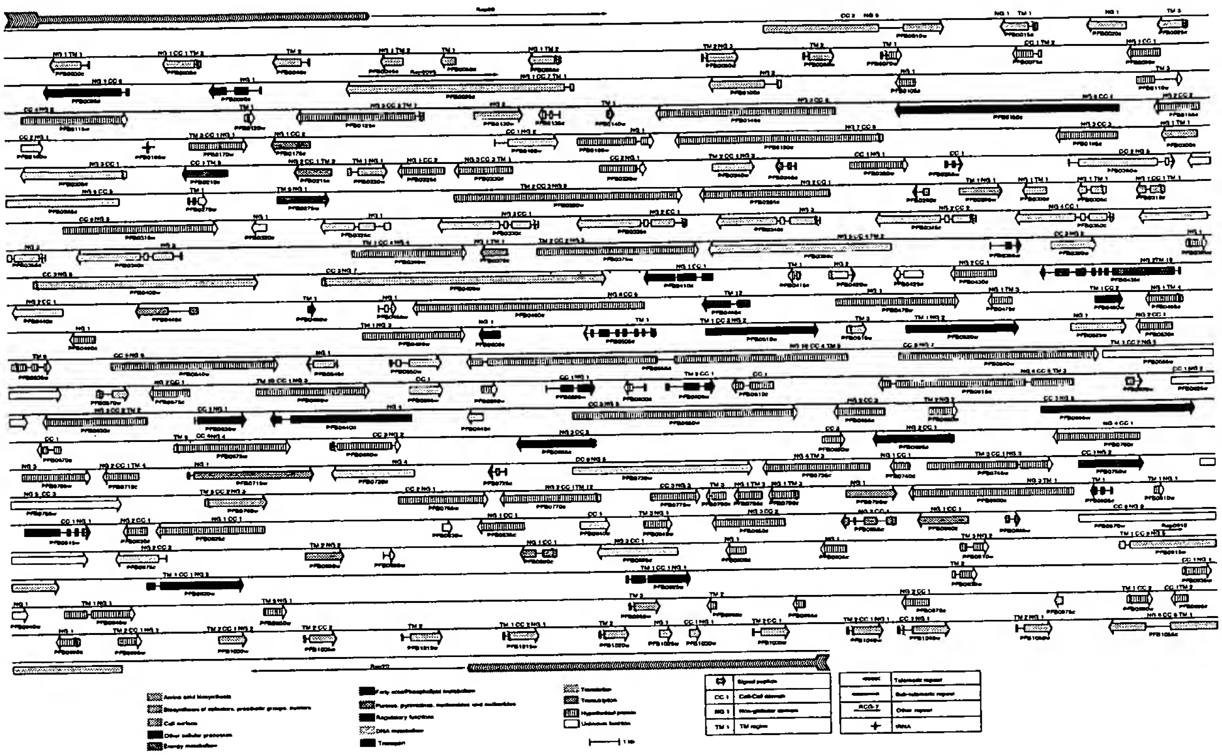

Fig. 1. Gene map of *P. falciparum* chromosome 2. Predicted coding regions are shown on each strand. Exons of protein-encoding genes are indicated by rectangles, and lines linking rectangles represent introns. The single tRNA^{Glu} gene is indicated by a cloverleaf structure. Genes are color-coded according to broad role categories as shown in the key. Genes identification numbers correspond to PF numbers in Table 2. The letters CC, NG, and TM followed by numerals indicate the number of predicted coiled-coil, nonglobular, and transmembrane domains in the proteins, respectively.

because some introns may have been missed by the gene-finding method. Most spliced genes consist of two or three exons. In terms of intron content and gene density, the *Plasmodium* genome, which was assessed by the analysis of the first completed chromosome sequence, appears to be intermediate between the condensed yeast genome and the intron-rich genomes of multicellular eukaryotes.

The proteins encoded in chromosome 2 (Table 2) fall into the following three categories: (i) 72 proteins (34%) are conserved in other genera and contain one or more distinct globular domains; (ii) 47 proteins (23%) belong to *Plasmodium*-specific families with identifiable structural features and, in some cases, known functions; and (iii) 90 predicted proteins (43%) have no detectable homologs, although many contain structural features such as signal peptides and transmembrane domains. Homologs outside *Plasmodium* were detected for 87 (42%) of the 209 predicted proteins. These include proteins in the first category, in addition to those proteins in the second category that possess a conserved domain or domains that are arranged in a manner unique to *Plasmodium*. The percentage of evolutionarily conserved proteins is about two times lower than that found for other geñomes, mainly because most of the remaining proteins were predicted to consist primarily of nonglobular domains (*15*) (Table 1). The abundance of nonglobular domains in *Plasmodium* proteins is very unusual; the proportion of proteins with predicted large nonglobular domains in other eukaryotes, such as *S. cerevisiae* (Table 1) or *C. elegans* (*16*), is approximately half that observed in *Plasmodium*. Furthermore, 13 of the 87 conserved proteins on chromosome 2 appear to contain large nonglobular structures (>30 amino acids) that are inserted directly into globular domains, as determined by alignment with homologs from other species.

To determine whether nonglobular domains and proteins are expressed in *P. falciparum*, we performed a reverse transcriptase (RT)–PCR on 11 nonglobular domains and on two genes that encoded predominantly nonglobular proteins, using total blood-stage RNA as a template. In all cases, RT-PCR products were the same size as those that were amplified from genomic DNA, and the sequence of RT-PCR products matched the genomic DNA sequence (*17*). Thus, it is likely that most, if not all, predicted nonglobular domains in chromosome 2 genes are expressed. One example of the insertion of a nonglobular domain into a well-defined globular domain is seen in a protein containing a 5′-3′ exonuclease (Fig. 2). The alignment of the *Plasmodium* sequence with four bacterial exonucleases revealed a 176–amino acid insertion in a region between a strand and a helix in the three-dimensional structure of this protein (*18*). This suggests that eukaryotic proteins can accommodate inserts that may be excluded from the protein core folding without impairing the protein function. The propagation of nonglobular domains in *Plasmodium* suggests that such proteins provide specific selective advantages to the parasite. A structural analysis of *Plasmodium* proteins that contain nonglobular inserts may be valuable for understanding the general principles of protein folding.

Of the 87 conserved proteins that are encoded on chromosome 2, 71 (83%) show the greatest similarity to eukaryotic homologs (Table 2). In contrast, the remaining 16 proteins are most similar to bacterial proteins, and 4 of these represent the first eukaryotic members of protein families that have previously been seen only in bacteria. At least some of these 16 genes may have been transferred to the nuclear genome from an organellar genome after the divergence of the phylum Apicomplexa from other eukaryotic lineages. Several of these proteins appear to contain NH_2-terminal organellar import peptides (*19*) and may function within the apicoplast or the mitochondrion. One such gene encodes 3-ketoacyl–acyl carrier protein (ACP) synthase III (FabH), which catalyzes the condensation of acetyl–coenzyme A and malonyl-ACP in type II (dissociated) fatty acid synthase systems. Type II synthase systems are restricted to bacteria and the plastids of plants, confirming previous hypotheses that the *Plasmodium* apicoplast contains metabolic pathways that are distinct from those of the host (*20*, *21*).

Because the phylum Apicomplexa represents a deep branch in the eukaryotic tree, the presence of eukaryotic-specific genes in *P. falciparum* suggests the appearance of these genes early in eukaryotic evolution. Most of these genes code for proteins that are involved in DNA replication, repair, transcription, or translation (Table 2) and include the origin recognition complex subunit 5, excision repair proteins ERCC1 and RAD2, and proteins involved in chromatin dynamics (such as the BRAHMA helicase, an ortholog of the DRING protein containing the RING finger domain, and chromatin protein SNW1). Furthermore, several eukaryotic proteins involved in secretion are encoded in chromosome 2 (such as the SEC61 γ subunit, the coated pit coatamer subunit, and syntaxin), suggesting an early emergence of the eukaryotic secretory system.

Proteins of the DnaJ superfamily act as cofactors for HSP70-type molecular chaperones and participate in protein folding and trafficking, complex assembly, organelle biogenesis, and initiation of translation (*22*). Five proteins containing DnaJ domains are present on chromosome 2, which suggests multiple roles for this domain in the *Plasmodium* life cycle. Two of these proteins consist primarily of the DnaJ domain, whereas three of the five proteins also contain a large nonglobular domain. Several proteins containing a DnaJ domain have been detected on other chromosomes, indicating that this is a large gene family in *Plasmodium* (*23*). One of its members, the ring-infected erythrocyte surface antigen, binds to the cytoplasmic side of the erythrocyte membrane, suggesting that DnaJ domains perform chaperone-like functions in the formation of protein complexes at this location (*24*). DnaJ domains in some *P.*

Table 1. Summary of features of *P. falciparum* chromosome 2 (*P. f.* chr 2) and comparison to *S. cerevisiae* chromosome 3 (*S. c.* chr 3). Protein structural features were predicted as described (*14*). ND, not determined. Numbers in parentheses indicate the percentage of the total genes or proteins with the specified properties.

Description	Number	
	P. f. chr 2	*S. c.* chr 3
Chromosome length (kb)	945	315
Percent G+C content	19.7	38.6
Exons	24.3	40.0
Introns	13.3	ND
Kilobases per gene	4.50	1.73
Number of predicted protein-coding regions	209	171
Number of genes with introns (%)	90 (43)	4 (2.2)
tRNA genes	1	10
Class of proteins		
Total	209	171
Secreted (%)	22 (11)	11 (6)
Integral membrane (%)	90 (43)	42 (24)
Integral membrane with multiple predicted transmembrane domains (%)	27 (13)	21 (12)
Containing coiled-coil domains (%)	111 (53)	32 (19)
Containing other large compositionally biased regions with predicted nonglobular structure (%)	155 (74)	71 (41)
Completely nonglobular (%)	17 (8)	6 (3.5)
With detectable homologs in other species	87 (42)	145 (85)

Table 2. Identification of genes on *P. falciparum* chromosome 2. The PF number is the systematic name assigned according to a method adapted from *S. cerevisiae* (*14*). The description contains the name (if known) and prominent features of the gene. The table includes genes with homologs in other species and members of *Plasmodium* gene families. An expanded version of this table with additional information is available on the World Wide Web at www.tigr.org/tab/mdb/pfdb/pfdb.html. Prt, protein; OO, organellar origin; TP, transit peptide; ATP, adenosine triphosphate; euk., eukaryotic; nt, nucleotide.

PF number	Description
Amino acid biosynthesis	
PFB0200c	Aspartate aminotransferase
Biosynthesis of cofactors, prosthetic groups, and carriers	
PFB0130w	Prenyl transferase
PFB0220w	Ubiquinone biosynthesis methyltransferase
Fatty acid and phospholipid metabolism	
PFB0385w	Acyl-carrier prt
PFB0410c	Phospholipase A2-like a/b fold hydrolase
PFB0505c	3-ketoacyl carrier prt synthase III, FabH (OO, TP)
PFB0685c	ATP-dependent acyl-CoA synthetase (TP)
PFB0695c	ATP-dependent acyl-CoA synthetase (TP)
Purines, pyrimidines, nucleosides, and nucleotides	
PFB0295w	Adenylosuccinate lyase (OO)
DNA metabolism	
PFB0160w	ERCC1-like excision repair prt
PFB0180w	Prt with 5′-3′ exonuclease domain (OO, TP)
PFB0205c	Prt with 5′-3′ exonuclease domain (Kem-1 family)
PFB0265c	RAD2 endonuclease
PFB0440c	Chromatinic RING finger prt, DRING ortholog
PFB0720c	Origin recognition complex subunit 5 (ATPase)
PFB0730w	BRAHMA ortholog (DNA helicase superfamily II)
PFB0840w	Replication factor C, 40-kDa subunit (replication activator)
PFB0875c	Chromatin-binding prt (SKI/SNW family)
PFB0895c	Replication factor C, 140-kDa subunit (ATPase)
Energy metabolism	
PFB0795w	ATP synthase alpha chain
PFB0880w	FAD-dependent oxidoreductase (OO)
Transcription	
PFB0140w	Metal-binding prt (DHHC domain)
PFB0175c	Prt of the MAK16 family
PFB0215c	Prt with Egl-like 3′-5′ exonuclease domain
PFB0245c	RNA polymerase 16-kD subunit, RPB4-like
PFB0255w	RRM-type RNA-binding prt
PFB0290c	Zn-ribbon transcription factor (TFIIS family)
PFB0370c	RNA-binding prt (KH domain)
PFB0445c	eIF-4A–like DEAD family RNA helicase
PFB0620w	YOU2-like small euk. C2C2 Zn finger prt
PFB0715w	DNA-directed RNA polymerase subunit 2
PFB0725c	Meta-binding prt (DHHC domain)
PFB0855c	rRNA methylase (SpoU family) (OO, TP)
PFB0860c	RNA helicase
PFB0865w	Small nuclear ribonucleoprt. (SNRNP family)
PFB0890c	Pseudouridine synthetase (RsuA family); first euk. member (OO)
Translation and post-translational modification	
PFB0165w	tRNA-Glu
PFB0240w	PINT domain prt (proteasomal subunit)
PFB0260w	PSD2-like 26S proteasomal subunit
PFB0325c	SERA antigen/protease with active Cys
PFB0330c	SERA antigen/protease with active Cys
PFB0335c	SERA antigen/protease with active Cys
PFB0340c	SERA antigen/protease with active Ser
PFB0345c	SERA antigen/protease with active Ser
PFB0350c	SERA antigen/protease with active Ser
PFB0355c	SERA antigen/protease with active Ser
PFB0360c	SERA antigen/protease with active Ser
PFB0380c	phosphatase (acid phosphatase family)
PFB0390w	Ribosome releasing factor (OO, TP)
PFB0455w	Ribosomal prt L37A
PFB0515w	Glycosyl transferase (novel euk. family)
PFB0525w	Asparaginyl-tRNA synthetase (OO, TP)
PFB0545c	Ribosomal prt L7/L 12 (OO)
PFB0550w	Euk. peptide chain release factor
PFB0585w	Leu/Phe-tRNA prt transferase, first euk. member (OO)
PFB0645c	Ribosomal prt L13 (OO)
PFB0830w	Ribosomal prt S26
PFB0885w	Ribosomal prt S30
Regulatory functions	
PFB0150c	Ser/Thr prt kinase
PFB0510w	GAF domain prt (cyclic nt signal transduction)
PFB0520w	Novel prt kinase
PFB0605w	Ser/Thr prt kinase
PFB0665w	Ser/Thr prt kinase
PFB0815w	Calcium-dependent prt kinase (C-terminus EF hand)
Transport	
PFB0210c	Monosaccharide transporter
PFB0275w	Membrane transporter
PFB0435c	Predicted amino transporter
PFB0465c	Membrane transporter
Cell surface	
PFB0010w	*var* gene
PFB0015c	Rifin
PFB0020c	*var* gene fragment
PFB0025c	Rifin
PFB0030c	Rifin
PFB0035c	Rifin
PFB0040c	Rifin
PFB0045c	*var* gene fragment
PFB0050c	Rifin pseudogene
PFB0055c	Rifin
PFB0060w	Rifin
PFB0065w	Rifin
PFB0100c	Knob-associated His-rich prt
PFB0300c	Merozoite surface antigen MSP-2
PFB0305c	Merozoite surface antigen MSP-5 (EGF domain)
PFB0310c	Merozoite surface antigen MSP-4 (EGF domain)
PFB0400w	PfS230 paralog (predicted secreted prt)
PFB0405w	Transmission-blocking target antigen PfS230
PFB0570w	Predicted secreted prt (thrombospondin domain)
PFB0760w	Mtn3/RAG1IP-like prt
PFB0915w	RESA-H3 antigen
PFB0955w	Rifin
PFB0975c	*var* gene fragment
PFB1000w	Rifin pseudogene
PFB1005w	Rifin
PFB1010w	Rifin
PFB1015w	Rifin
PFB1020w	Rifin
PFB1025w	*var* gene fragment
PFB1030w	*var* gene fragment
PFB1035w	Rifin
PFB1040w	Rifin
PFB1045w	*var* gene fragment
PFB1050w	Rifin
PFB1055c	*var* gene
Other cellular processes	
PFB0085c	Prt with DnaJ domain (RESA-like)
PFB0090c	Prt with DnaJ domain
PFB0450w	Prt translocation complex, SEC61 γ chain
PFB0480w	Syntaxin
PFB0500c	RAB GTPase
PFB0595w	Prt with DnaJ domain, DNJ1/SIS1 family
PFB0635w	T-complex prt 1 (HSP60 fold superfamily)
PFB0640c	WEB-1 ortholog, WD40
PFB0750w	VPS45-like prt (STXBP/UNC-18/SEC1 family)
PFB0805c	Clathrin coat assembly prt
PFB0920w	Prt with DnaJ domain (RESA-like)
PFB0925w	Prt with DnaJ domain (RESA-like)
Unknown function	
PFB0270w	SLR1419 family prt (OO)
PFB0320c	HesB family prt (possible redox activity, OO, TP)
PFB0420w	YgdB prt first euk. member (OO, TP)
PFB0425c	YMR7 family prt

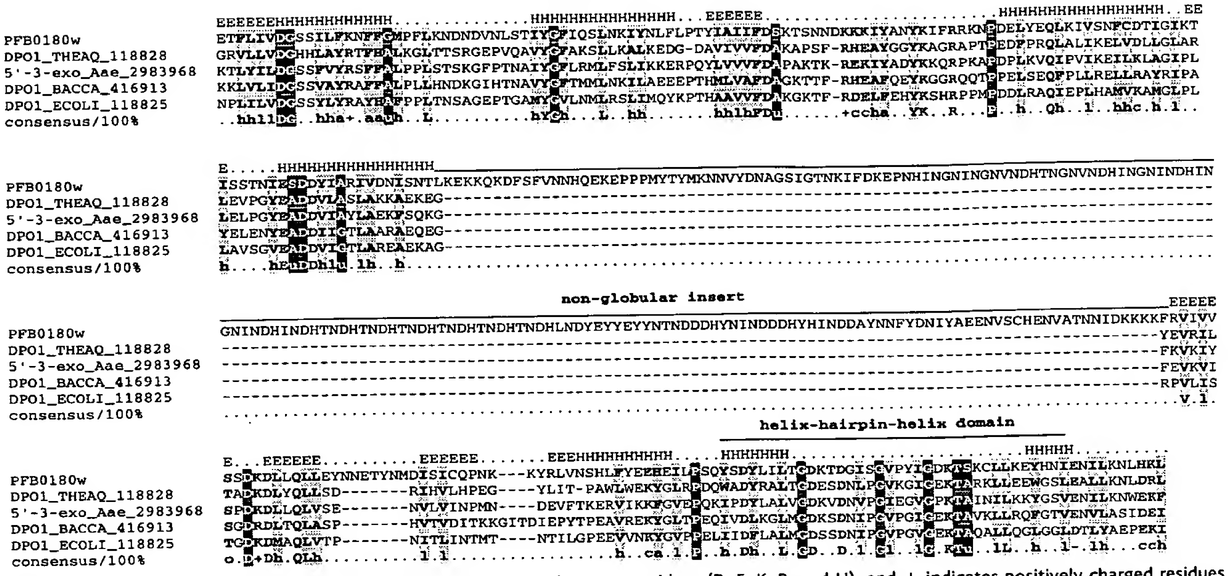

Fig. 2. Multiple alignment of the predicted 5'-3' exonuclease (PFB0180w) encoded in chromosome 2 with homologous bacterial exonuclease domains showing the large nonglobular insert in *Plasmodium*. The alignment was constructed with the profile alignment option of CLUSTALW (*34*). The alignment column shading is based on a 100% consensus, which is shown underneath the alignment; h indicates hydrophobic residues (A, C, F, I, L, M, V, W, and Y), u indicates "tiny" residues (G, A, and S), o indicates hydroxy residues (S and T), c indicates charged residues (D, E, K, R, and H), and + indicates positively charged residues (K and R) (*35*). The aspartates involved in metal coordination have a red background and inverse type. Secondary structure elements derived from the crystal structure of *Thermus aquaticus* DNA polymerase (*18*) are shown above the alignment (H indicates α helix, and E indicates extended conformation, or β strand). 5'-3'-exo_Aae is a stand-alone exonuclease from *Aquifex aeolicus*, and the remaining bacterial sequences are the NH_2-terminal domains of DNA polymerase I.

falciparum proteins contain substitutions in the His-Pro-Asp signature that is required for interaction with HSP-70–type proteins, which may indicate a modification of the typical chaperone function.

Chromosome 2 contains five protein families that are unique to *Plasmodium* in terms of their distinct domain organization, although three of them contain domains that are conserved in other genera. The genes encoding the *Plasmodium*-specific families are primarily located near the ends of the chromosome. A single *var* gene was identified in each subtelomeric region. The *var* genes encode large transmembrane proteins (PfEMP1) expressed in knobs on the surface of schizont-infected red cells. PfEMP1 proteins exhibit extensive sequence diversity; are clonally variant; and are involved in antigenic variation, cytoadherence, and rosetting (*6–8*). In addition to the full-length *var* genes, six small ORFs were identified in the subtelomeric regions that were similar to *var* sequences. Five of these ORFs resembled the *var* exon II cDNAs or the Pf60.1 sequences that were reported previously (*7*, *25*).

The largest *Plasmodium*-specific family found on chromosome 2 encodes proteins that were dubbed rifins, after the RIF-1 repetitive element. RIF-1 contained a 1-kb ORF but no initiation codon, was found on most chromosomes, and was transcribed in late blood-stage parasites (*9*). The function of the RIF-1 element was unknown. Eighteen ORFs with similarities to RIF-1 were found in the subtelomeric regions of chromosome 2, centromeric to the *var* genes. An inspection of the sequence upstream of these ORFs revealed exons encoding signal peptides, which indicated that the RIF-1 elements were actually genes consisting of two exons. These genes encode potential transmembrane proteins of 27 to 35 kD, with an extracellular domain that contains conserved Cys residues that might participate in disulfide bonding, a transmembrane segment, and a short basic COOH-terminus. The extracellular domain also contains a highly variable region (Fig. 3). RT-PCR with schizont RNA showed that one of six rifin genes that were tested was transcribed. The function of the rifins is unknown, but their sequence diversity, predicted cell surface localization, and expression in schizont stages suggest that, like *var* genes, they may be clonally-variant. Multiple rifin genes were detected in the telomeric regions of chromosomes 3 and 14, suggesting that rifin genes have propagated as clusters in the course of *Plasmodium* evolution (*26*). If the number found on chromosome 2 is representative of other chromosomes, there may be 500 or more rifin genes in the *P. falciparum* genome (~7% of all protein-coding genes), making it the most abundant gene family in this organism. The presence of *var* and rifin genes and other ORFs in subtelomeric regions of *P. falciparum* chromosomes confirms that the subtelomeric regions are not transcriptionally silent (*27*).

Another family of membrane-associated proteins, serine repeat antigens (SERAs), contains a papain protease-like domain. A cluster of three SERA genes, which were all transcribed in the same direction (from centromere to telomere), was known to be on chromosome 2 (*28*); at least one SERA has been evaluated for use in blood-stage vaccines. These genes are part of an eight-gene cluster; seven genes have a similar four-exon structure, but the gene at the 3' end of the cluster contains only three exons. The protease domains in these proteins are unusual because five of the eight contain serine instead of cysteine in the active nucleophile position, suggesting that they are serine proteases with a structure that is typical of cysteine proteases (*29*).

Two proteins (MSP-4 and MSP-5) that contain an epidermal growth factor (EGF) module in their extracellular domains were identified (*30*, *31*). In organisms that are not classified in the animal kingdom, MSP-4,

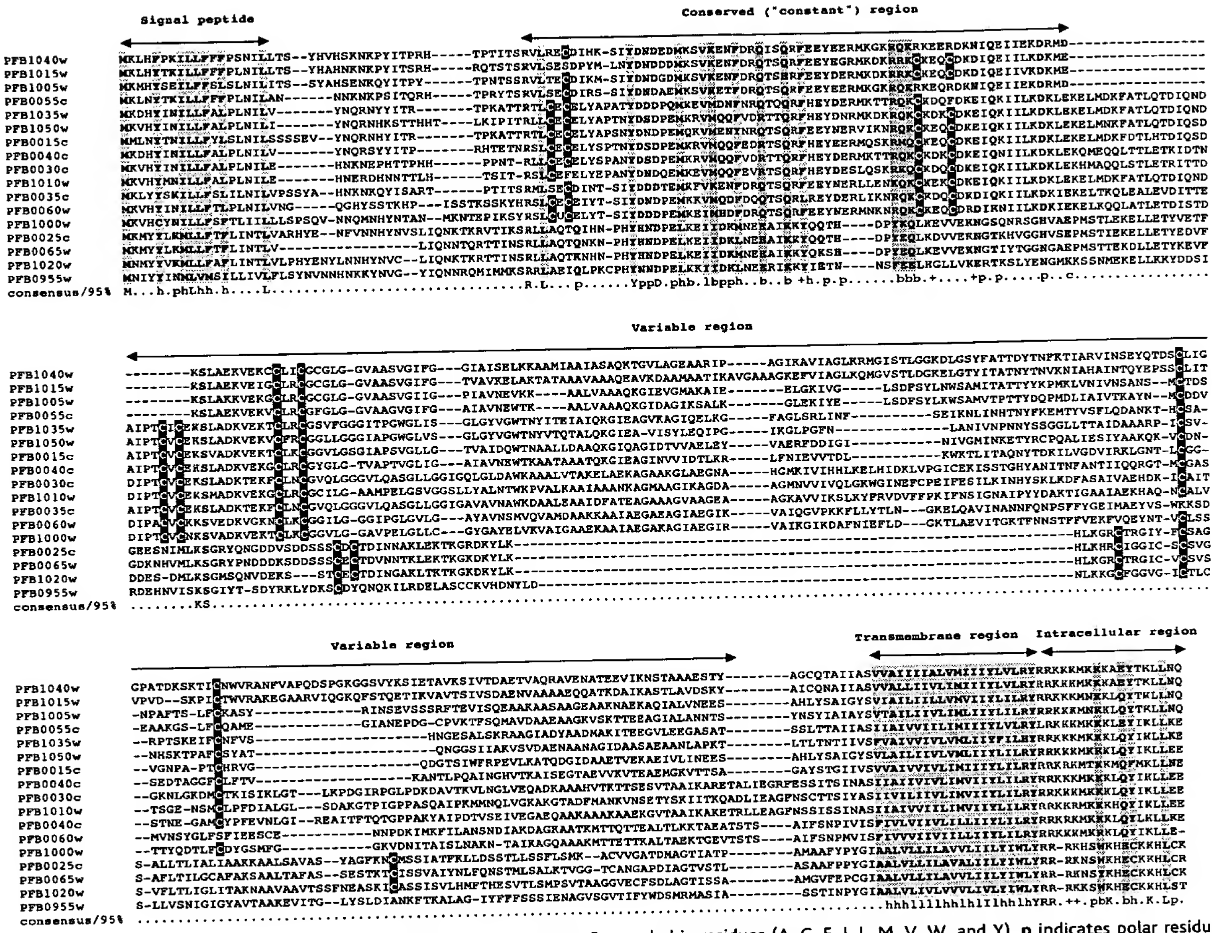

Fig. 3. Multiple sequence alignment of rifins encoded on chromosome 2. The predicted coding regions were aligned with CLUSTALW (*34*) using the default settings. The alignment column shading is based on a 95% consensus, which is shown underneath the alignment; **h** indicates hydrophobic residues (A, C, F, I, L, M, V, W, and Y), **p** indicates polar residues (D, E, H, K, N, Q, R, S, and T), **b** indicates "big" residues (F, I, L, M, V, W, Y, K, R, Q, and E), and **+** indicates positively charged residues (K and R) (*35*). The cysteines conserved in subsets of rifins are shown by inverse type.

MSP-5, and MSP-1 (a multi-EGF domain protein encoded on chromosome 3) and two *Plasmodium* sexual-stage antigens (*32*) are the only proteins that contain EGF repeats, which suggests that *Plasmodium* obtained the sequence for this domain from its animal host. The plasmodial EGF domains may be involved in parasite adhesion to host cells.

In addition to the families of *Plasmodium*-specific proteins, chromosome 2 contains genes for many secreted and membrane proteins. One of these genes encodes a protein with a modified thrombospondin domain and was transcribed in blood-stage parasites (*17*). Other *Plasmodium* proteins containing thrombospondin domains, such as sporozoite surface protein 2/TRAP and circumsporozoite protein, are involved in the parasitic invasion of host cells (*33*), suggesting that this protein may be involved in the binding of infected red cells to host-cell ligands.

Determination of the first *P. falciparum* chromosome sequence demonstrates that the A+T richness of *P. falciparum* DNA will not prevent the sequencing of the genome. Although technical difficulties not observed during the sequencing of other microbial genomes were encountered, solutions to these problems were found that will facilitate sequencing of the remaining chromosomes. The genome sequence should be of value in the study of *Plasmodium* biology and in the development of new drugs and vaccines for the treatment and prevention of malaria. In addition to these practical benefits, the *Plasmodium* genome sequence should provide broader biological insights, particularly in regard to the plasticity of the eukaryotic genome that is manifest in the preponderance of the predicted nonglobular domains in plasmodial proteins.

References and Notes

1. World Health Organization, *Wkly. Epidemiol. Rec.* **72**, 269 (1997).
2. J. B. Dame *et al.*, *Mol. Biochem. Parasitol.* **79**, 1 (1996); M. Lanzer, D. de Bruin, J. V. Ravetch, *Nature* **361**, 654 (1993); K. Suplick, R. Akella, A. Saul, A. B. Vaidya, *Mol. Biochem. Parasitol.* **30**, 289 (1988); M. J. Gardner, D. H. Williamson, R. J. M. Wilson, *ibid.* **44**, 115 (1991); R. J. M. Wilson *et al.*, *J. Mol. Biol.* **261**, 155 (1996); S. Köhler *et al.*, *Science* **275**, 1485 (1997).
3. R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995).
4. C. M. Fraser *et al.*, *ibid.* **270**, 397 (1995); C. J. Bult *et al.*, *ibid.* **273**, 1058 (1996); C. M. Fraser *et al.*, *Nature* **390**, 580 (1997); J.-F. Tomb *et al.*, *ibid.* **388**, 539

(1997); H. P. Klenk *et al.*, *ibid.* **390**, 364 (1997); C. M. Fraser *et al.*, *Science* **281**, 375 (1998).

5. *P. falciparum* clone 3D7 was selected because it can complete all stages of the life cycle and because 3D7 was used in a genetic cross [D. Walliker *et al.*, *Science* **236**, 1661 (1987)] and in The Wellcome Trust Malaria Genome Mapping Project [J. Foster, J. Thompson, *Parasitol. Today* **11**, 1 (1995)]. Parasites were grown in vitro [W. Trager and W. Jensen, *Nature* **273**, 621 (1978)] and embedded in agarose [D. J. Kemp *et al.*, *ibid.* **315**, 347 (1985)]. Chromosomes were resolved on preparative pulsed-field gels (the process used 1.2% SeaPlaque GTG agarose, a Bio-Rad DRIII apparatus, a 180- to 250-s switch time, a 120° field angle, and 3.7 V/cm for 90 hours at 14°C). Chromosome 2 bands from five gels were adjusted to 0.3 M sodium acetate to prevent melting of the AT-rich DNA and were digested with agarase. The exposure of the DNA to UV light was minimized. A shotgun library of 1- to 2-kb fragments was prepared in pUC18 as described (*3*), except that treatment with *Escherichia coli* DNA polymerase I was performed (0.5 mM deoxynucleoside triphosphates at 37°C for 10 min) after the second ligation step to close nicks before electroporation into DH10B cells. The gel-purified chromosome 2 DNA was only ~85% pure because of the co-migration of sheared DNA from other chromosomes. To compensate for this ~85% purity and to provide excess coverage to compensate for the possible nonrandomness of the shotgun library, we obtained 23,768 sequences (a coverage of about 10-fold). FS+ dye-terminator chemistry (Perkin-Elmer Applied Biosystems, Foster City, CA) was superior to dye-primer chemistry for the sequencing of AT-rich DNA. Sequences were assembled with The Institute for Genomic Research (TIGR) Assembler [G. S. Sutton, O. White, M. D. Adams, A. R. Kerlavage, *Genome Sci. Tech.* **1**, 9 (1995)], which was modified to assemble A+T-rich sequences. Neighboring contigs were identified with the program GROUPER (A. D. Mays, TIGR, Rockville, MD), and 10 groups of 114 contigs were mapped on the chromosome by comparison to sequence-tagged site (STS) markers [M. Lanzer, D. de Bruin, J. V. Ravetch, *Nature* **361**, 654 (1993)]. The closure of physical and sequence gaps was performed as described (*3*). Physical gaps were closed by PCR reactions with a genomic DNA template with primers from adjacent mapped groups or with primers from one mapped group and each of the unmapped groups. PCR reactions (Expand Long Template PCR System, Boehringer Mannheim) contained 100 ng of genomic DNA and 15 pmol of each primer (BioServe Biotechnologies, Laurel, MD) in a 50-ml reaction. Cycling conditions (Perkin-Elmer GeneAmp PCR Systems 9600 or 9700) were as follows: 94°C for 2 min; 10 cycles at 94°C for 1 min, at 50° or 55°C for 1 min, and at 60°C for 2 min; 20 cycles at 94°C for 1 min, at 50° or 55°C for 1 min, and at 60°C for 2 min plus 20 s per cycle; and 1 cycle at 60°C for 10 min. PCR products were purified (QIAquick PCR Purification Kit; QIAGEN, Chatsworth, CA) and sequenced with dye-terminator chemistry. Sequence gaps that were too rich in A+T for primer synthesis and walking were closed by the insertion of the artificial transposon AT-2 [S. E. Devine and J. D. Boeke, *Nucleic Acids Res.* **22**, 3765 (1994)] into the plasmid templates that spanned each sequence gap; multiple transposon-containing subclones of each template were sequenced to close the gaps. The coverage criteria were that every position required at least double-clone coverage (or sequence from a PCR product that was amplified from genomic DNA) and either sequence from both strands or coverage with two different sequencing chemistries. The sequence was edited manually with TIGR Editor, and additional sequencing reactions were performed to improve coverage and to resolve sequence ambiguities. To independently confirm the colinearity of the assembled sequence and genomic DNA, we prepared Nhe I and Bam HI optical restriction maps of chromosome 2 DNA [J. Jing *et al.*, in preparation] and compared them with restriction maps that were predicted from the sequence. The relative errors of predicted and observed fragment sizes were 4.3 and 5.8% for the Nhe I and Bam HI maps, respectively, indicating that the assembled sequence was an accurate representation of the chromosome. Further proof of colinearity was obtained by a comparison of the sequence to a scaffold of YAC-end sequences from chromosome 2 YACs that were isolated from a library provided by K. Hinterberg [J. Foster and J. Thompson, *Parasitol. Today* **11**, 1 (1995); L. Cummings *et al.*, in preparation].
6. D. I. Baruch *et al.*, *Cell* **82**, 77 (1995).
7. Z. Su *et al.*, *ibid.*, p. 89.
8. J. D. Smith *et al.*, *ibid.*, p. 101 (1995); J. A. Rowe, J. M. Moulds, C. I. Newbold, L. H. Miller, *Nature* **388**, 292 (1997).
9. J. L. Weber, *Mol. Biochem. Parasitol.* **29**, 117 (1988).
10. K. D. Vernick and T. F. McCutchan, *ibid.* **28**, 85 (1988).
11. P. Oquendo *et al.*, *ibid.* **18**, 89 (1986); J. Patarapotikul and G. Langsley, *Nucleic Acids Res.* **16**, 4331 (1988).
12. S. Saitoh, K. Takahashi, M. Yanagida, *Cell* **90**, 131 (1997); M. M. Smith *et al.*, *Mol. Cell Biol.* **16**, 1017 (1996); M. M. Mahtani and H. F. Willard, *Genome Res.* **8**, 100 (1998); R. D. Shelby, O. Vafa, K. F. Sullivan, *J. Cell Biol.* **136**, 501 (1997); D. du Sart *et al.*, *Nature Genet.* **16**, 144 (1997).
13. J. Lechner and J. Ortiz, *FEBS Lett.* **389**, 70 (1996); A. A. Hyman and P. K. Sorger, *Annu. Rev. Cell Dev. Biol.* **11**, 471 (1995).
14. The nonredundant (NR) protein sequence database at the National Center for Biotechnology Information (NCBI) (NIH, Bethesda, MD) was searched with the gapped BLAST and PSI-BLAST programs. Coding regions were predicted with GlimmerM, a eukaryotic gene-finding program based on Glimmer [S. L. Salzberg, A. L. Delcher, S. Kasif, O. White, *Nucleic Acids Res.* **26**, 544 (1998)], trained on a set of 117 *P. falciparum* sequences. Gene models based on GlimmerM predictions, similarity of ORFs to known proteins, and prediction of putative signal peptides and transmembrane domains were constructed with ANNOTATOR (L. Xhou, TIGR). In cases where a putative gene had no database match and multiple GlimmerM predictions of gene structure, the highest scoring model was reported. After the first set of models was inspected, it was added to the training set, and GlimmerM was retrained. Gene models should be regarded as preliminary until confirmed by other methods. Protein structural features were delineated with the UniPred program of the SEALS package [D. R. Walker and E. V. Koonin, *Ismb* **5**, 333 (1997)]. Signal peptides were predicted with SignalP [H. Nielsen, J. Engelbrecht, S. Brunack, G. von Heijne, *Protein Eng.* **10**, 1 (1997)], and transmembrane helices were predicted with PHDhtm [B. Rost, R. Casadio, P. Fariselli, C. Sander, *Protein Sci.* **4**, 521 (1995)]. Coiled-coil domains were predicted with COILS (J. Kuzio, NCBI). Nonglobular structures were predicted with SEG [J. C. Wooton and S. Federhen, *Methods Enzymol.* **266**, 554 (1996)]. Multiple sequence alignments were constructed with CLUSTALW or with the Gibbs-sampling option of the MACAW program [G. D. Schuler, S. F. Altschul, D. J. Lipman, *Proteins* **9**, 180 (1991); A. F. Neuwald, J. S. Liu, C. E. Lawrence, *Protein Sci.* **4**,1618 (1995)]. Transfer RNAs were identified with tRNAscan [T. M. Lowe and S. R. Eddy, *Nucleic Acids Res.* **25**, 955 (1997)]. Systematic gene names based on a scheme for *S. cerevisiae* [H. W. Mewes *et al.*, *Nature* **387** (suppl.), 7 (1997)] were assigned with the convention PF (for *P. falciparum*), a letter for the chromosome (A for chromosome 1, B for chromosome 2, and so forth), a three-digit code ordering the genes from left to right in increments of five (to allow for the addition of new genes), and a letter denoting the coding strand (w or c, for Watson or Crick strand, respectively).
15. The term "nonglobular" refers to proteins or domains of proteins that do not assume compact, folded structures [J. C. Wootton, *Comput. Chem.* **18**, 269 (1994)]. There is a strong inverse correlation between compositional bias in protein sequences and their ability to fold into a compact, globular domain [J. C. Wootton and S. Federhen, *Methods Enzymol.* **266**, 554 (1996)]. Accordingly, the compositional complexity of a sequence can be used to partition it into predicted globular and nonglobular domains. In this analysis, the prediction was performed with the SEG program with the following parameters: window length, 45; trigger complexity, 3.4; and extension complexity, 3.75.
16. L. Aravind and E. Koonin, unpublished data.
17. D. J. Carucci *et al.*, data not shown.
18. Y. Kim *et al.*, *Nature* **376**, 612 (1995).
19. V. Haucke and G. Schatz, *Trends Cell Biol.* **7**, 103 (1997).
20. A. R. Slabas and T. Fawcett, *Plant Mol. Biol.* **19**, 169 (1992); R. J. M. Wilson, M. J. Gardner, J. E. Feagin, D. H. Williamson, *Parasitol. Today* **7**, 134 (1991).
21. After this manuscript was submitted for publication, we learned of work that confirmed the identification of the 3-ketoacyl-ACP synthase III gene in *Plasmodium* and the importation of nuclear-encoded proteins into the apicoplast in the related parasite *Toxoplasma* [R. F. Waller *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 12352 (1998)].
22. D. M. Cyr, T. Langer, M. G. Douglas, *Trends Biochem Sci.* **19**, 176 (1994).
23. L. Aravind *et al.*, data not shown.
24. P. Bork, C. Sander, A. Valencia, B. Bukau, *Trends Biochem. Sci.* **17**, 129 (1992); J. Watanabe, *Mol. Biochem. Parasitol.* **88**, 253 (1997); R. L. Coppel *et al.*, *Nature* **310**, 789 (1984); I. A. Quakyi *et al.*, *Infect. Immun.* **57**, 833 (1989); M. Foley, L. Corcoran, L. Tilley, R. Anders, *Exp. Parasitol.* **79**, 340 (1994).
25. S. Bonnefoy, E. Bischoff, M. Guillotte, O. Mercereau-Puijalon, *Mol. Biochem. Parasitol.* **87**, 1 (1997).
26. Sequence data for *P. falciparum* chromosome 3 was obtained from the Sanger Centre (available at http://www.sanger.ac.uk/Projects/P_falciparum/). Sequencing of *P. falciparum* chromosome 3 was accomplished as part of the Malaria Genome Project Consortium with support by the Wellcome Trust.
27. R. R. Hernandez *et al.*, *Mol. Cell Biol.* **17**, 604 (1997); K. Fischer *et al.*, *ibid.*, p. 3679 (1997).
28. B. Knapp, E. Hundt, U. Nau, H. A. Kupper, *Mol. Biochem. Parasitol.* **32**, 73 (1989); B. Knapp, U. Nau, E. Hundt, H. A. Kupper, *ibid.* **44**, 1 (1991); W. B. Li, D. J. Bzik, T. Horii, J. Inselburg, *ibid.* **33**, 13 (1989); B. A. Fox and D. J. Bzik, *ibid.* **68**, 133 (1994).
29. D. G. Higgins, D. J. McConnell, P. M. Sharp, *Nature* **340**, 604 (1989); A. E. Eakin, J. M. Higaki, J. H. McKerrow, C. S. Craik, *ibid.*, **342**, 132 (1989).
30. V. M. Marshall *et al.*, *Infect. Immun.* **65**, 4460 (1997).
31. V. M. Marshall, W. Tieqiao, R. L. Coppel, *Mol. Biochem. Parasitol.* **94**, 13 (1998).
32. L. Aravind, unpublished observations; M. J. Blackman, I. T. Ling, S. C. Nicholls, A. A. Holder, *Mol. Biochem. Parasitol.* **49**, 29 (1991); D. C. Kaslow *et al.*, *Nature* **333**, 74 (1988); P. E. Duffy, P. Pimenta, D. C. Kaslow, *J. Exp. Med.* **177**, 505 (1993).
33. K. J. Robson *et al.*, *Nature* **335**, 79 (1988); C. Cerami *et al.*, *Cell* **70**, 1021 (1992); W. O. Rogers *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 9176 (1992).
34. J. D. Thompson, D. G. Higgins, T. J. Gibson, *Nucleic Acids Res.* **22**, 4673 (1994).
35. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

36. We thank the members of the Malaria Genome Sequencing Consortium for the open discussion of data during the development of the effort to sequence the *P. falciparum* genome; D. J. Lipman and L. H. Miller for helpful discussions; M. Gottlieb for support and encouragement; A. Craig for providing the 3D7 clone and for suggestions on pulsed-field gel electrophoresis; P. de la Vega for the culturing of parasites; M. Lanzer for providing STS data; K. Hinterberg for providing the 3D7 YAC library; and the TIGR faculty, sequencing core, bioinformatics staff, and systems administrators for expert advice and assistance. This work was supported by a supplement to the National Institute of Allergy and Infectious Diseases grant R01 AI40125-01, which was made possible with funds from NIH's Office for Research on Minority Health; Department of the Army Cooperative Agreement grant DAMD17-98-2-8005 (to J.C.V.); and Naval Medical Research and Development Command Work Units 61102A.S13.00101.BFX1431, 612787A.870.00101.EFX.1432, 623002A.810.00101.HFX.1433, and STEP C611-102A0101BCX. The opinions and assertions herein are those of the authors and are not to be construed as official or as reflecting the views of the U.S. Navy or Department of the Army.

29 June 1998; accepted 29 September 1998